

UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S  
THESIS

---

# Validation of White Matter Hyperintensities Automatic Segmentation Methods

---

*Author:*  
Àlex ARCAS CUERDA

*Supervisor:*  
Dr. Eloi PUERTAS & Dr.  
Joaquim RADUÀ

*A thesis submitted in partial fulfillment of the requirements  
for the degree of MSc in Fundamental Principles of Data Science*

*in the*

Facultat de Matemàtiques i Informàtica

June 30, 2020



UNIVERSITAT DE BARCELONA

# *Abstract*

Facultat de Matemàtiques i Informàtica

Fundamentals of Data Science Master's Degree

## **Validation of White Matter Hyperintensities Automatic Segmentation Methods**

by Àlex ARCAS CUERDA

This master's thesis seeks to review and objectively evaluate the current white matter hyperintensities (WMH) automatic segmentation methods published journals. To this end, the methods have been systematically searched in scientific databases, and those meeting inclusion criteria have been evaluated. The evaluation has consisted in applying the method to detect WMH in our dataset of patients with bipolar disorder and healthy controls, in which an experienced neuroradiologist had manually coded all WMH.

After the systematic search, we selected all available methods that were ready for use with standard MRI data by a standard user. Four methods met these criteria. We then applied these methods to detect WMH in our dataset, and compared the results with the neuroradiologist-based ground truth deriving several evaluation metrics. This master's thesis also include a discussion section, in which we compare the results of our evaluations with the results of the WMH Segmentation Challenge held in 2017, which included substantially different datasets.

The most relevant conclusion of this master's thesis is that no method seems to be accurate enough for clinical implementation, although the low performance of the methods may be related to the differences between our data and the data that were used to train them. Besides, realizing the huge improvement made in the field during the last few years after the appearance of deep neural networks, we anticipate that a method with sufficient accuracy might be available soon.

The codes used to obtain the results and graphs displayed in this project together with some guidelines to run them are available through [PFM-WMH](https://github.com/aarcascuerda/PFM-WMH)<sup>1</sup>.

---

<sup>1</sup><https://github.com/aarcascuerda/PFM-WMH>



## *Acknowledgements*

I would like to thank the following people, without whom it would have been much harder to complete this project.

Special thanks to my supervisor Dr. Eloi Puertas, for his support as an expert in Data Science as well as his help to go through the whole process of developing this research in such difficult times. And special thanks to Dr. Joaquim Raduà, whose insight and knowledge into the subject steered me through this research. They both gave me great ideas which made this work considerably better than I could have thought.

The main organizer of the White Matter Hyperintensities Challenge, Hugo J. Kuijf, who helped me through the whole process of preprocessing my data through an endless series of mails which he had no obligation to reply. He did, along with clear answer to my misunderstandings. He also handed their data which helped a lot to test our methods and compare results although those couldn't be published.

The main developer of Coroflo, Robin Caramassa, who provided me the whole source code for the Coroflo algorithm plus helped me over some issues we had while building the method.

My colleagues at the masters degree with whom I have spend so many hours during this year. Specially this last semester in which I have only complained about how hard it was to be doing a PFM while confined.

And my biggest thanks to my family for all the support you have shown me through this research, the culmination of many years learning science.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Systematic Review</b>	<b>3</b>
2.1 Search Strategy . . . . .	3
2.2 Special Mentions . . . . .	4
2.3 The White Matter Hyperintensities Segmentation Challenge . . . . .	6
2.3.1 PGS . . . . .	7
2.3.2 Coroflo . . . . .	8
2.3.3 NeuroML 2 . . . . .	8
2.3.4 BigRBrain 2 . . . . .	9
2.3.5 Discarded methods . . . . .	9
<b>3 The Data</b>	<b>11</b>
3.1 Original Data . . . . .	11
3.1.1 3DT1 and Binary Face Mask . . . . .	11
3.1.2 FLAIR . . . . .	11
3.1.3 Aligned T1 and Transformation Parameters . . . . .	12
3.2 Preprocessed Data . . . . .	13
3.3 Manual Segmentation of White Matter Hyperintensities . . . . .	13
3.3.1 WMH, distributions . . . . .	14
3.3.2 Manual Reference Standard Masks . . . . .	14
<b>4 Evaluation</b>	<b>17</b>
4.1 Evaluation Metrics . . . . .	17
4.1.1 Dice Similarity Coefficient (DSC) . . . . .	17
4.1.2 Hausdorff Distance . . . . .	18
4.1.3 Average Volume Difference . . . . .	19
4.1.4 Recall and F1-score for individual lesions . . . . .	19
4.1.5 Final Rank . . . . .	20
4.2 Evaluation Results . . . . .	21
4.2.1 Dice Similarity Coefficient (DSC) . . . . .	21
4.2.2 Hausdorff Distance . . . . .	22
4.2.3 Average Volume Difference . . . . .	23
4.2.4 Recall for individual lesions . . . . .	24
4.2.5 F1-score for individual lesions . . . . .	26
4.2.6 Final Rank . . . . .	26
<b>5 Final Discussion</b>	<b>29</b>
<b>6 Further Work</b>	<b>31</b>

<b>A Systematic Review Table</b>	<b>33</b>
<b>Bibliography</b>	<b>43</b>



# 1 Introduction

During the last decades, neuroscience has rapidly risen its popularity among the scientific world. With it, many research groups readressed their efforts into the field. Not only focusing on the study of the nervous system as an anatomical or physiological system but also using day to day advances in molecular biology, chemistry or electrophysiology for example. Moreover, advances in further related sciences as mathematics, computational science or physics, have also proved to be useful. Some are currently being used to develop complex models of the brain which can get a better understanding of the processes underlying certain aspects of cognition. With all these new perspectives, we can now categorize the modern neuroscience as a multidisciplinary science.

One of the current topics of interest in neuroscience is the study of white matter hyperintensities (WMH). White matter hyperintensities are white matter areas that show abnormally high peaks of intensity in magnetic resonance imaging (MRI). They are usually presumed of vascular origin and are usually associated with regions of the brain which suffer from reduced blood flow. Micro-strokes or micro-hemorrhages are some of the proposed causes for them [1][2][3][4][5].

White matter hyperintensities are known to be common among elder people but they are found in young adults too. It has also been shown that these abnormalities in the white matter tissue of the brain are correlated with some mental disorders as dementia. Furthermore, recent studies show that subjects with deep white matter hyperintensities are up to three times more likely to suffer from bipolar disorder and major depressive disorder [6][7]. Considering these findings as well as other related with the causes or effects of WMH, interest in doing research on these abnormalities of the white matter tissue has arisen.

To study the WMH of a subject, one has first to locate them. This is usually accomplished through a series of MRI sequences of the subject. To do so, the typical MRI sequences used is the Fluid-attenuated inversion recovery (FLAIR) sequence, because with it WMH are easier to identify [2]. An example of a FLAIR sequence and a standard high-resolution T1 sequence are shown in Fig.1.1. Other type of sequences as for example proton density sequence (PD) had also been used to help identifying WMH. These other kinds of sequences are not that common through the clinical world and they do not supply much information about WMH, and thus they are currently being deprecated for this purpose.

Segmenting white matter hyperintensities in MRI sequences has become a key point as many recent studies depend on it. The major drawback is the process of segmenting WMH by hand, which is extremely time consuming and needs from an expert in the field. Moreover, there is no gold standard for this segmentation and many discrepancies appear when combining manual segmentation of different experts. To overcome these problems, many research groups are focusing on developing an automatic segmentation procedure.

Following this path of work, many automatic segmentation methods have been proposed over the last two decades but, as shown by the systematic review of automatic segmentation methods for WMH done in 2015 by Calguiri et al., Ref.[8], there

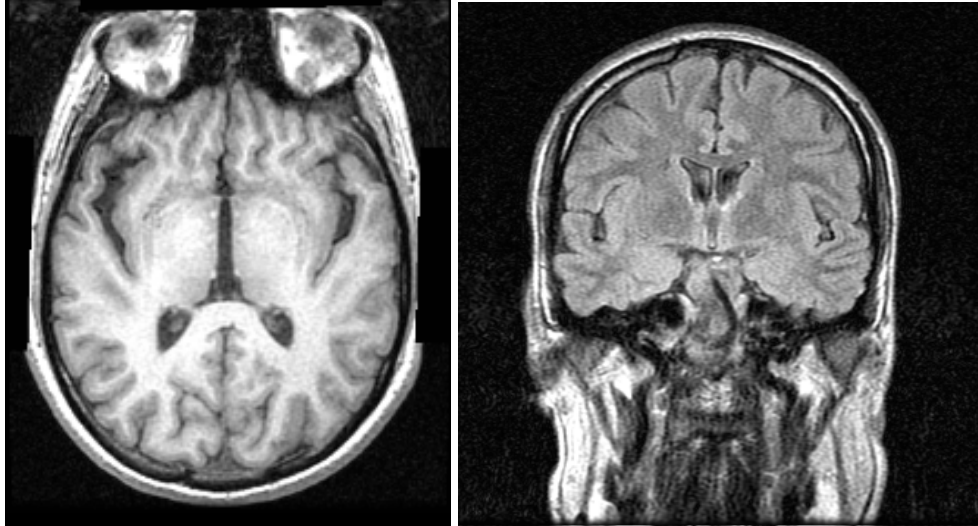


FIGURE 1.1: Examples of brain MRI sequences for our subjects. **Left:** Slice of T1-weighted sequence. **Right:** Slice of FLAIR sequence.

are significant issues when comparing these methods. First, common evaluation procedure is lacking, especially because no gold standard is available for manual segmentation. Second, many of the methods depend on the data that their authors used to train them. Most methods are too specific for certain MRI sequences or even for certain MRI configurations.

The first aim of the present master's thesis is to update the work by Calguiri et al. This is important because the progress made in the last 5 years has been huge. As an example, most of the methods evaluated by Calguiri et al. did not use deep learning algorithms. At present time, the standard has completely switched and most new methods use them. In fact, the fight for best scores has moved to which network architecture fits better to solve the issue of WMH segmentation. To understand some of the methods commented in this project some basic knowledge of deep learning would be useful, we recommend a lecture like Ref.[9].

The results of the systematic search as well as the actual selection criteria are discussed in Sec.2. While doing the Systematic Review, the only work fulfilling all the selection criteria was a challenge started at 2017 but currently up for evaluation [10]. The details of the challenge along with a brief description of the different methods used is addressed in Sec.2.3. The second aim of the present master's thesis is to objectively evaluate the methods that could be ready implemented in clinical setting. To this end, we compare the WMH that they detect in our dataset of patients with bipolar disorder and healthy controls, with the WMH that were carefully manually coded by an experienced neuroradiologist. The data used for the present work is described in Sec.3. The preprocessing steps applied to our data as well as some statistical analysis of it is also displayed in this section. Then, at the core of the work is Sec.4 where the metrics used for evaluation are described first. After, the performance of each method is presented together with a small report on of it. Finally, in Sec.5 all the results are discussed to give a sense of closure to the whole project. In it, the different issues and strengths of the current situation for WMH segmentation are explained. Sec.6 is committed to give some ideas for posterior projects on the field.

## 2 Systematic Review

With the purpose of updating the list of methods available in the literature, this project started by doing a systematic review of white matter hyperintensities automatic segmentation methods. A systematic review is a review study of available methods for a certain topic. The key idea behind a systematic review is that the review must detail the search strategy followed. This search strategy regularly involves a series of rules adopted when deciding to drop or accept one of the data sources, methods in our case. This procedure is best determined according to given standards such as the ones in [11]. The main idea behind these studies is to be transparent and easily reproducible. This methodology fits the current study perfectly as it will be shown that, although there are many papers centered on WMH automatic detection, they are not always easily applicable or even reproducible. Many reasons play into it but they will be discussed later through the project.

### 2.1 Search Strategy

The search was performed on PubMed<sup>1</sup> which is a freely available search motor. It searches in MEDLINE<sup>2</sup> database which is mainly related to articles and publications for health related sciences. Since vocabulary on the topic is quite variable, the search tried to account for the different naming alternatives for the automatic segmentation of white matter hyperintensities. The candidate publications were selected on February 29, 2020. The search term was chosen after carefully trying different variants. Then, finding the exact term which best fitted our search purposes. The exact search term is presented below.

**PubMed Search Term.** (*“white matter hyperintensity” OR “white matter hyperintensities” OR WMH*) AND (*detection OR segmentation*) AND (*automated OR semiautomated OR automatic OR semiautomatic*)

The search returned 128 candidate publications from which 80 were directly dropped as they were not related to the issue of interest. Because they did not present any automatic or semiautomatic method to segment WMH or they were an evaluation of algorithms or they used algorithms from other articles on the search. This was determined after either reading the abstract of the articles or by reading the appropriate sections referring to the algorithm used in the article. After dismissing these 80 publications, 48 remained. These were examined to find its respective segmentation algorithm. During this process, a clear appreciation was made, most of them just described an algorithm, some in detail, others partially. Then, the following selection criteria was to select those papers in which the algorithm source code was publicly available. All those publications in which the code had to be implemented, trained and tested were dropped. In this cut, 13 remained and 35 were dropped. From those which remained, the ways of distributing the codes were

---

<sup>1</sup><https://www.ncbi.nlm.nih.gov/pubmed>

<sup>2</sup><https://medlineplus.gov>

divers. Some claimed to deliver the code by email when requested while others had their code published in their websites or public repositories like GitHub or Docker-Hub.

Finally, 13 publications passed through our selection criteria. Then, one last selection criteria was used to select through those. The code for the model had to be trained, if possible, with a general perspective. This criteria may seem really strict but it is possibly the key point for our study. The whole purpose of the study, and we hope that also of the segmentation models, is to aid neuroscientists doing research on WMH. In this path, it is not wise to assume that the collective would know and have the time or motivation to learn about machine learning and how to train an algorithm. This is not an easy topic, even less when referring to deep learning and some of the mathematical concepts behind it. We think the process should be the other way around : data scientists should try to give a *ready-to-use* method, as this is their field of expertise. In the ideal scenario, neuroscientists would just input their MRI sequences and obtain WMH location masks as an output, hopefully, with good accuracy. This may sound as a too high shoot right now but it will arrive sooner or later. In fact, we prove this scenario not to be so far ahead.

Therefore, using our last selection criteria, only the article of the White Matter Hyperintensities Challenge survived, see Ref.[10]. As the methods on this challenge are our final choice for the evaluation, Sec.2.3 is fully devoted to explain it in detail.

Finally, all the selection criteria used for the present study are summed up below:

- I Search in PubMed using the reference term described in 2.1.
- II Drop all those publications which did not include an algorithm for automatic or semiautomatic segmentation of WMH. In case they repeated one, the original publication describing the algorithm was chosen.
- III Drop all those publications which did not hand a way to get the source code of their methods or an application to use them.
- IV Drop all those publications which do have the source code for the segmentation model but do not provide the training of the model. Hence, the model has to be previously trained and fitted before being operative.

As a remark, all 128 publications and their dropping reasons are detailed in Appendix A.

## 2.2 Special Mentions

Although most methods in the search did not fulfil the selection criteria, there are some which may be appropriate for special cases. Those methods pertain all to the group of the 13 methods which passed the first three selection criteria. Methods whose code was not even implemented would definitely need an expert on the field plus, in most cases, a considerable amount of time devoted to code, train and test them. Then, in this section we refer to those methods which did not pass the selection criteria but do have certain relevance.

### BIANCA

Brain Intensity AbNormality Classification Algorithm [12], or BIANCA for short, is the WMH segmentation algorithm that comes along with the FSL package since v.6

(current version). The creators of BIANCA themselves claim the algorithm to be in a BETA state and advice to verify all the results extracted by the algorithm before using them.

Nevertheless, BIANCA comes along with FSL which is one of the most used packages in neuroimaging. Hence it is worth mentioning it just for that. BIANCA is based on a k-NN algorithm that classifies MRI voxels depending on their intensity and spatial features. The major drawback of BIANCA is that it strongly depends on the set of parameters chosen for training (directly stated by the owners of the method). Then, if one understand how to properly train and tune the algorithm to the desired data, it should obtain good performance. A strong point of BIANCA is its great flexibility in terms of MRI modalities, it is prepared to face all the modalities if provided with a training dataset with the same modalities.

Worth note to mention that, although the process of training the algorithm and tuning the parameters of the model is not trivial, the owners provide a user guide on their [website](#)<sup>3</sup> to facilitate the task.

## CASCADE

CASCADE [13] is an open-source software package based on FSL with the finality of *defining* WMH. This *learned* definition can then be used to segment WMH on a dataset. Technically speaking, the algorithm is a machine learning algorithm but also a purely statistical procedure. The work develops a statistical test based on the intensity distribution of the image to assess the probability of a voxel to be a WMH. Then, the parameters of the algorithm are tuned on data before being able to output results.

The performance of the method doesn't score high in any of the reviews we have gone through but the fact that it actually encloses the definition of a WMH, which is not yet clear, makes it worth mentioning.

## U-NET

Before U-Net [14] arrived, most of the methods for automatic segmentation of WMH were based on k-NN, SVM or Bayesian algorithms. The biggest problem that deep learning faced when used for segmentation of brain regions was the lack of data. Usually, big neural networks rely on thousands of samples for the train stage. In the case of MRI sequences, having access to that number of samples is usually not feasible due to privacy rights.

In May 2015, with the purpose of avoiding the lack of huge biomedical imaging sets for training, U-Net was born. In our research it was clear that from that point, most of the new methods rely on deep learning methodologies. A high percentage of them being adaptations of U-Net to WMH segmentation. As an example, two of the 13 independent finalists use U-Net as base architecture for their neural networks [15][16]. In our review, the original paper for U-Net [14] did not appear as the algorithm was created for general biomedical segmentation purposes. Their creators developed it to introduce a deep learning architecture that could face biomedical imaging segmentation with a small number of training samples. To do so, they created this *U* architecture which heavily depended on data augmentation and on segmentation of the images at different scales. As seen in Fig. 2.1, U-Net is basically a set of convolutional layers downsampling the image and then upsampling it to the initial shape where the fully connected layers perform the classification. The whole

<sup>3</sup><https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA>

point of it is that each *scale* of the convolutional layers not only connects with the one of bigger and lower dimensionality but with the one with the same dimensionality through a *bridge* connection. Nowadays, U-Net has proved its effectiveness not only in segmenting WMH but many other tissues too.

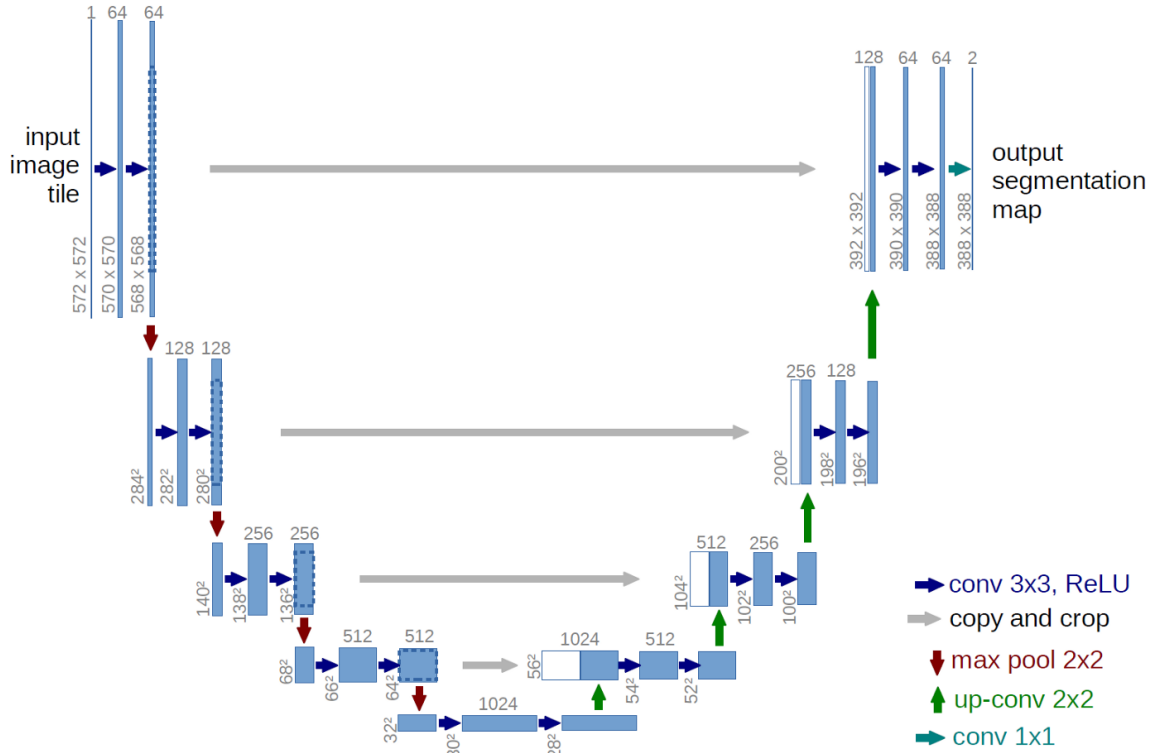


FIGURE 2.1: U-Net original architecture.

## 2.3 The White Matter Hyperintensities Segmentation Challenge

The White Matter Hyperintensities Segmentation Challenge, WMH Segmentation Challenge from now on, is a competition challenge organized as a joint effort of the UMC Utrecht, VU Amsterdam and NUHS Singapore. The main reference for the work is Ref.[10] and further information can be found in their [website](#)<sup>4</sup>.

The main point for choosing this article is that as mentioned in Sec.2.1, all methods submitted to the challenge fulfil the selection criteria of our study (not some of those which were submitted after the challenge final date). In principle, all the methods can be freely pulled from DockerHub to be directly used for segmentation without previous training or debugging. In fact, we think the goal of our project to be extremely related to theirs. Finding and training an algorithm for a general perspective is a must as, new studies may use different scanners or even different imaging protocols and thus, citing Ref.[10], *many (deep) machine learning methods require some form of transfer learning or fine-tuning on the target images, which in practice is not always feasible*.

The official challenge was announced at Quebec on 2017 and the ending was on 2019. Despite, as the real purpose of it was to provide an standardized evaluation

<sup>4</sup><https://wmh.isi.uu.nl>



platform for WMH segmentation methods, methods can still be uploaded and will be evaluated and classified along with the rest of the methods. Actually, since the publication of the official results in November 2019, more than 10 new segmentation methods have been submitted. Four of them ranking Top 5 and two of them even outperforming the original challenge winner.

At 1st of April 2020, more than 35 teams had submitted their methods for evaluation in the challenge. This project just focuses on those methods ranked Top 10. Next, these methods are described, the models they are based on together with the preprocessing or postprocessing steps they performed. Understand that the exact process to train these models is not described as it is out of the scope of this project. The full description of each algorithm and performance is available through “[results](#)”<sup>5</sup>.

### 2.3.1 PGS

PGS [17] is currently the best ranked method of the competition. The method first preprocesses the images using ROBEX [18] to extract the non-brain regions of the FLAIR and T1 images. Then, it does a gaussian normalization of all the intensities in between the lower 5% and upper 95%. Finally, it crops all the images with 200x200 patches in the axial plane (this is a key point in its performance with our data, further discussed in Sec.4).

The base model for the method is on the deep learning framework being U-Net its basic building block. Its structure consists on a U-Net with convolutions of 3x3 or 5x5, batch normalization, max pooling, ELU activation function and a final convolution layer 1x1 plus softmax for the output probabilities.

The model runs five times randomly choosing different initialization and batch. Then, the segmentation results are aggregated and thresholded to find the final segmentation result.

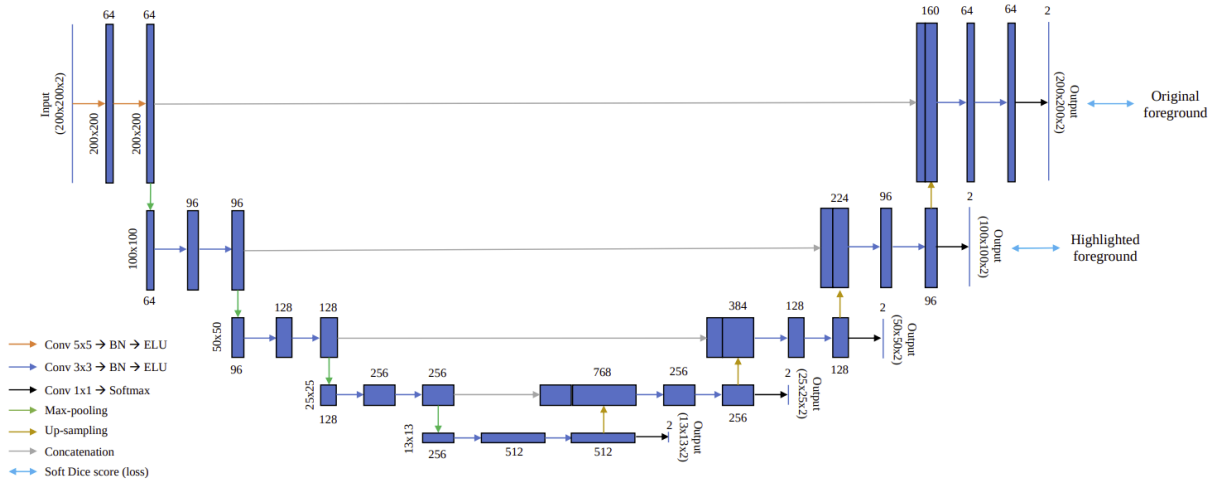


FIGURE 2.2: PGS version of U-Net architecture. Image from Ref.[17].

Note that this method had a maximum resolution of 500 voxels per axis, as we have images bigger than that, we had to fix this. To do so, we went through the code to adapt it to our data resolution. Still, the model is not trained or thought for our image resolution so it may not perform as expected.

<sup>5</sup><https://wmh.isi.uu.nl/results>

### 2.3.2 Coroflo

Coroflo [19] is currently ranked as the fifth method of the competition. The method first uses BET2 to extract the brain. The images are then split in tiles of  $70 \times 70 \times 22$  or  $72 \times 72 \times 24$  voxels which overlap a 50% with each other to get a robust output. Before fitting the image into the network, the method also does an intensity normalization. It takes those intensities with percentiles below 1% or above 99% and sets them to 0 or 1 respectively.

The model is based on MD-GRU layers linked with channel-wise fully connected layers. Gated recurrent units are an architecture of recurrent neural networks like LSTM but without an output gate, see Fig.2.3. The MD just stands for multi-dimensional as our data consists on 3D images. Further reading on GRU networks can be found in Ref.[20].

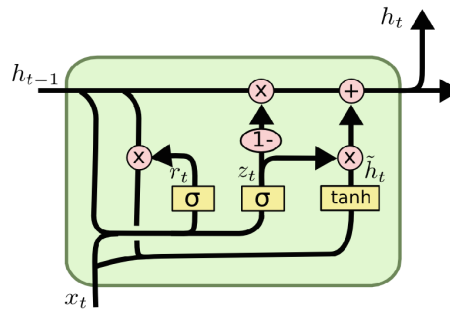


FIGURE 2.3: Scheme of GRU architecture where,  $x_t$  is the input vector,  $h_t$  is the output vector,  $r_t$  is the reset gate vector,  $z_t$  is the update gate vector,  $\tilde{h}_t$  is the candidate vector and  $h_{t-1}$  is the output vector of the previous iteration.

The model is applied over several spatial configurations before summing the outputs. Remark that the original MD-GRU is also used by Cian (the seventh ranked) whose authors were also the authors of the publication in which MD-GRU was first used for brain tissues segmentation [21].

After doing the prediction, Coroflo outputs a series of tiles which have to be reconstructed to a full size image. The final results are then averaged over 5 different configurations of the model and thresholded to 0 or 1.

At first, Coroflo did not work for our data although the authors claim it should. Consequently, we had to go through the code trying to adapt it to our data. During the process, we could just fix the two main configurations of the full model (5 in total). Although these two configurations are able of determining a high percentage of the WMH by themselves, being able to use the 5 different configurations would certainly improve the method performance.

### 2.3.3 NeuroML 2

NeuroML 2 [22] is currently ranked as the sixth method of the competition being the improved version of NeuroML which ranked on position thirty-six. The method firstly preprocesses the images by doing a high resolution brain extraction on the T1-weighted sequence with FSL-BET. Then, the method takes the parameters of the registration of both main sequences and applies it on the brain mask. Finally, it applies the brain mask to both T1 (registered, explained in Sec.3.1) and FLAIR images. To assure the performance of the brain extraction, the method erodes the slides at the



extreme of the vertical axes. Once the brain is cleanly extracted, the method scales the intensities by setting those one below 5% to 0 and those ones above 95% to 1.

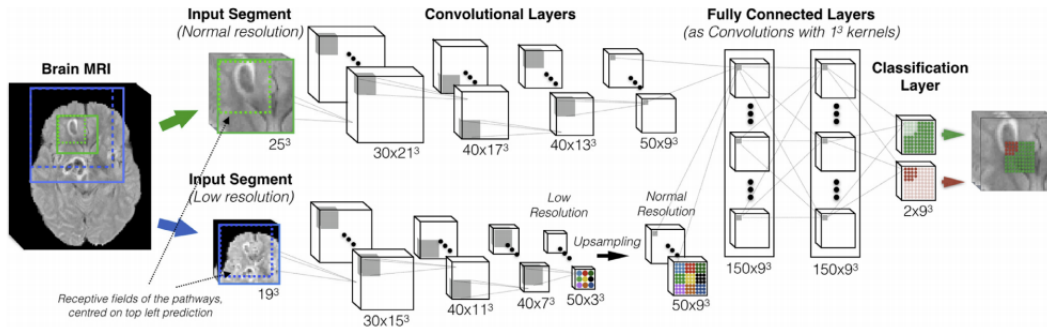


FIGURE 2.4: Example of a double-pathway DeepMedic architecture for multi-scale processing. At each layer, the number and size of feature maps is depicted in the format. Actual DeepMedic has 11 layers by default.

As a base model, the method uses a modification of the DeepMedic [23]. DeepMedic is network based on two main components, 3D convolutional layers and a 3D fully connected conditioned random field. The architecture is quite different from U-Net although it also uses the idea of multi-scale processing but through a double-pathway architecture, see Fig. 2.4. First modification the authors made of DeepMedic is a size expansion of the image incoming patches. Then, they also modified the network by adjusting the patch sampling technique as well as adjusting the ratio of patches with lesions.

To compute the final results, they aggregated over 5 different initialization of the above described model.

### 2.3.4 BigRBrain 2

BigRBrain 2 [24] is currently ranked as the tenth method of the competition being an improved version of BigRBrain which is ranked on position eleven. The method does not use any specific preprocessing and uses patches of size 200x200x16 as network input.

The model of the method is a clean U-Net using the dice coefficient as loss function (see Sec. 4.1.1). The method applies a custom probability map on the outputs of the network named by the authors as Posterior-CRF. The details of this conditional random field (CRF) reference map which is the key aspect of this method can be found in [24].

The model was trained on four different samples of the data to give the final results by aggregating the results of each trained model and thresholding over them.

### 2.3.5 Discarded methods

From the Top 10 methods only 4 were finally evaluated. From the rest of the methods, SysuMedia (the winner of the WMH Segmentation Challenge), SysuMedia 2 (an improvement of the first), NLP Logix and NIH Cidi 2 were discarded because they could not handle the high resolution of our data. Before actually dropping them, we revised the code trying to debug it or adapt it to allow images with higher resolution. Both SysuMedia proved data specific as the algorithm starts by checking the shape of the input image conditioning the following steps on this information (the

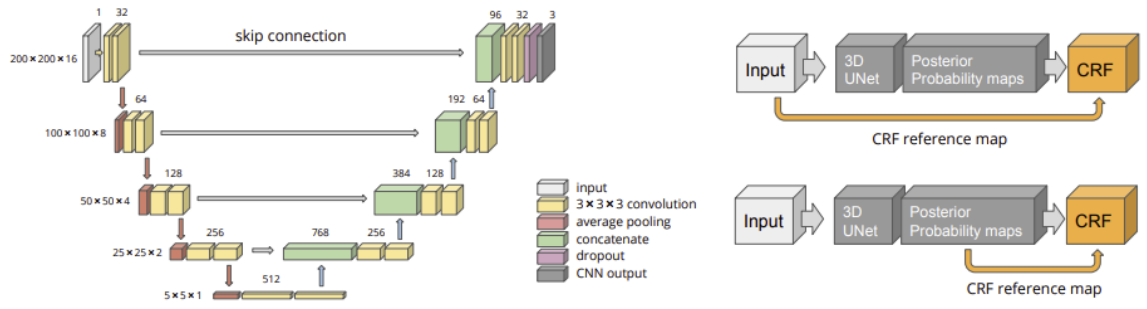


FIGURE 2.5: **End-to-end training networks.** For each graph: 3D UNet baseline (left), Intensity-CRF (upper right) and Posterior-CRF Neural Network (lower right). Image from Ref.[24]

algorithm basically expect the images to shape as in the challenge). This procedure may have proved useful to win the competition but it is not what was expected. Then for NLP Logix and NIH Cidi 2, the neural network model was not prepared to handle the shape of our images as input. We found no clear way of modifying these two methods to adapt them to our data.

Cian and Anonymous 20200413 were two methods submitted after the end of the competition just for evaluation purposes. Due to this, they do not need to make publicly available their codes and in fact, they didn't do so. We were quite interested in Cian as it uses MD-GRU architecture. This architecture has been performing better than U-Net in certain circumstances as it seems not to overfit as much. We contacted the developers of Cian during the realization of the present work but we did not get a reply. For Anonymous 20200413, there is not even a public email or author so there was no clear way of getting access to the method.

Discarding these methods we do not imply that they are of poor quality. As a matter of fact, they are among the ones with best performance in the challenge. However, they do not fit into our review as the whole purpose of it is to find and evaluate truly general and freely available methods. Take into consideration that, if one faces a problem in which the resolution of the MRI sequences is the expected by these methods, they are worth a try.

## 3 The Data

Initially, our data consisted on MRI sequences of 167 subjects and the manual segmentation their WMH by an expert neuroradiologist. From all those samples, we had to drop the data of 6 subjects due to incompleteness, either in the manual segmentation part or due to some of the MRI sequences being missing or corrupt. Then, the whole study was performed on 161 subjects, 67 suffered from bipolar disorder and 94 were patients with no mental disorder known.

This section starts by describing all the elements in the original data which is used directly by some of the segmentation methods. Secondly, the preprocessing steps and methodologies made on the data before feeding it on the segmentation methods is described. Finally, there is an explanation of how the WMH masks were created from the manual segmentation data. Some relevant aspects of their spatial and size distributions are also remarked as they are important for evaluation.

### 3.1 Original Data

The original data consists on a set of 5 files for each subject. The description of each file and how it was obtained is described in the following sections.

#### 3.1.1 3DT1 and Binary Face Mask

The 3DT1 is a Neuroimaging Technology Initiative or NifTI for short. It's an image corresponding to the T1-weighted MRI sequence in which the face has been removed. As the exact procedure to deface the 3DT1 image and obtain the binary mask was not specified in the WMH Segmentation Challenge, we chose to use the Deface module of the FSL library, see Ref.[25]. There are many other libraries to deface an MRI sequence but FSL is one of the standard packages. Then, using the Deface module, we extracted the binary mask for the face of each subject and the resultant T1 image with the face cropped. After doing so, we checked visually for each subject that the face extraction performed well and did not crop undesired parts.

The T1-weighted sequences were taken in the transverse direction ( $\hat{z}$ ), see Fig.3.1. The resulting image shape of the MRI sequence is  $(\hat{x}, \hat{y}, \hat{z}) \rightarrow (512, 512, 256)$ . The voxels on the 3DT1 are isotropic as their respective size is of 1 mm in each direction,  $1.0 \times 1.0 \times 1.0 \text{ mm}^3$ . Obviously, the image space of the binary face mask is equal to the one of the 3DT1.

#### 3.1.2 FLAIR

The FLAIR is a NifTI image corresponding to the fluid attenuated inversion recovery MRI sequence, this far, no preprocessing has been made to it. The manual segmentation was realized on this image as is where the WMH are visually appreciable. Then for our project and in general, the coordinates of the WMH are in this image space. As a consequence of this, the algorithms must run in this image space to be able to evaluate the results.

The FLAIR sequences were taken in the coronal direction ( $\hat{y}$ ), see Fig.3.1. The resulting image shape of the sequences is  $(\hat{x}, \hat{y}, \hat{z}) \rightarrow (512, 22, 512)$ . The pixels in the  $\hat{x}$  and  $\hat{z}$  directions correspond to 0.4297 mm each and the ones on the  $\hat{y}$  direction correspond to 6 mm each. Each voxel of the 3D image has a size of  $0.4297 \times 6.0 \times 0.4297 \text{ mm}^3$ . The FLAIR image space is relevant as is where the manual delimitation is performed and where the algorithms output their results for the WMH segmentation. Note that the FLAIR image space is strongly an-isotropic as the  $\hat{x}$  and  $\hat{z}$  directions have almost 14 times the resolution of the  $\hat{y}$  direction.

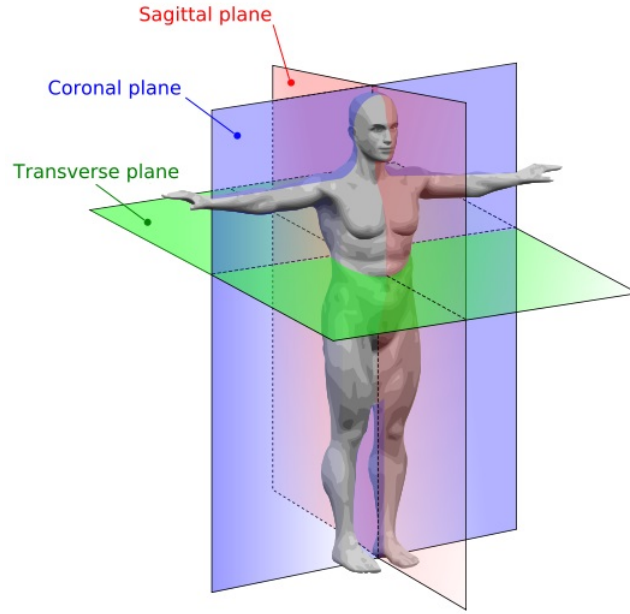


FIGURE 3.1: MRI sequence planes.

### 3.1.3 Aligned T1 and Transformation Parameters

The plain 3DT1 does not serve of much because our segmentation methods are built to work on the FLAIR image space. We need to align both images to be able to correlate each voxel in the FLAIR image to one or more of the 3DT1 image. This is a recurrent problem in neuroimaging and it's been widely studied. Many possible transformations exist, some pretty simple and other more complex. Actually, we previously disposed of a registration of both images but, it was performed with a different package than the one used in the WMH Segmentation Challenge, the FSL package. We wanted to replicate as accurately as possible the preprocessing of the WMH Segmentation Challenge so we did a new registration using the Elastix toolbox [26]. To do so, we adapted the parameters given in the challenge for our data when required. The transformation parameters are given in a text file along with the final T1 image aligned in the FLAIR space. From now on, we keep referring to the not aligned T1 images as 3DT1 and to the aligned T1 images simply as T1. In fact, this process of registration of both images could be thought as a preprocessing step. If we've done so is because some of the methods rely on it but most simply take the original 3DT1 image and do the registration with their own procedures.

## 3.2 Preprocessed Data

Apart from the *original* data, the challenge provided samples of the FLAIR, T1 and 3DT1 images with a bias field correction. This data is already considered preprocessed although the bias field correction is just one of the small and general steps before training a model to segment WMH. Despite the correction, most methods on the challenge perform further preprocessing on the data to achieve better performance results. The decision of whether to choose the original data or the bias field corrected one is on them. Some may even use both samples.

The bias field correction was performed with a MATLAB library named SPM12 [27]. The version used in the challenge has been deprecated hence we used a newer one. Nevertheless, the parameters used for the bias field correction were taken from the standards in the WMH Segmentation Challenge. The process was applied at the three full images, FLAIR, T1 and 3DT1. The outputs are smoothed versions of the original images, a sample of each is shown in Fig. 3.2.

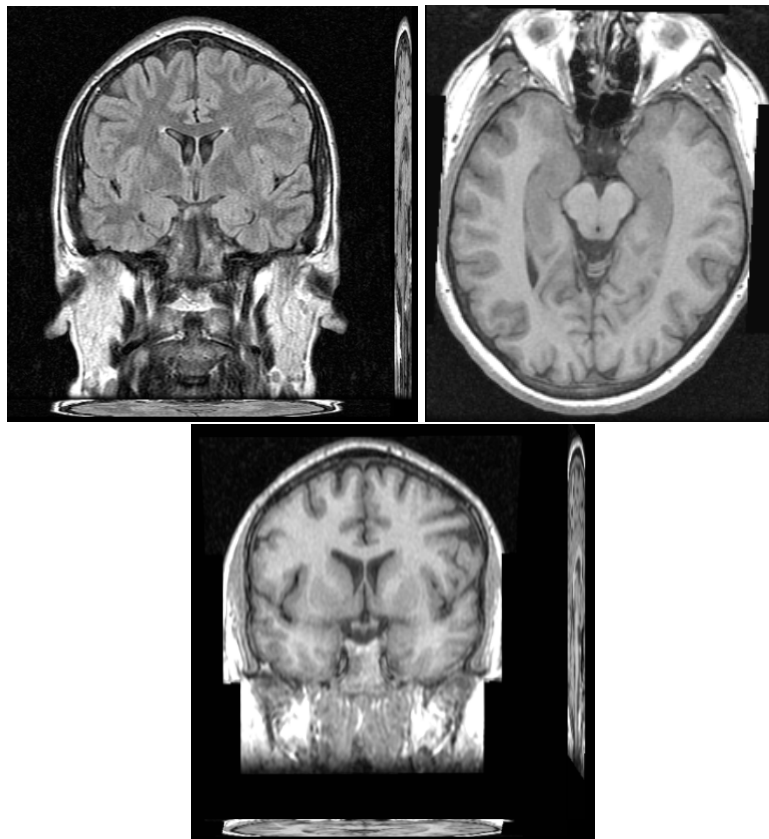


FIGURE 3.2: Bias field corrected sections of a FLAIR image (**Top-Left**), a 3DT1 image (**Top-Right**) and a T1 image (**Bottom**)

## 3.3 Manual Segmentation of White Matter Hyperintensities

Our initial data contained the manual segmentation of WMH for each subject. The exact data was formed by the central coordinates of each WMH, the subject it pertained and a label giving its approximate size. The labels were divided in four different size categories, *big* for WMH larger than 5 mm, *medium* for those WMH in between 5 and 2 mm, *small* for WMH shorter than 2 mm and *dot* for those WMH

which were just one voxel. All the annotations had been taken by a expert neuroradiologist and revised twice. From now on, we consider this data our manual standard as is the only reliable source we have available for the positions and sizes of the WMH in our data.

### 3.3.1 WMH, distributions

Before generating the evaluation masks and to give some insight about our data, we computed some statistics of our manual segmentation's. These are displayed and discussed below.

First, by checking the number of WMH present on each subject, we realized that all our subjects had at least 20 WMH independently of suffering or not dementia. In fact, acknowledging if a subject suffers from dementia can't be directly extracted from the WMH distribution as, Fig.3.3 left does not display two distinct regions. Deeply analyze patterns on the distributions, sizes and other characteristics of the WMH in brains suffering from dementia or other mental disorders is the key point for which automatic detection of WMH is important.

Another important realization was made while checking the distribution of sizes. In it, we found that a 92.33% were classified as *dot*, see Fig.3.3 right. This should be considered for further discussion while doing the evaluation because the *dot* classified hyperintensities were pointed out to be of questionable nature by the expert neuroradiologist.

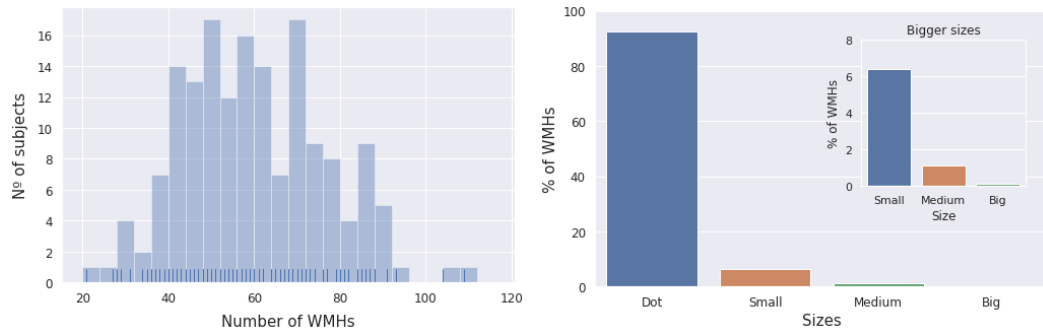


FIGURE 3.3: **Left:** Distribution of WMH in our subjects data. **Right:** Distribution of WMH sizes in the reference data. Both figures are done taking into account the manual segmentation of the expert neuroradiologist.

Despite all the subjects having at least 20 WMH, non of them had more than 120. Then the distribution had  $\mu = 59.54$  and  $\sigma = 16.48$ . We already realized that most of the WMH were *dot* size, hence less than 2 mm of radius at max. This means that one WMH is roughly 27 voxels of our image. Having images of 512x512x22 voxels means having a total of 5,767,168 voxels per image. Then, making a rough estimation we find that only 0.03% of the voxels in our images correspond to WMH. This being said, we state that our data is extremely unbalanced. Having data so unbalanced special care should be taken when defining the evaluation metrics to try to avoid redundancy in the results.

### 3.3.2 Manual Reference Standard Masks

As counterpart of our manual segmentation of WMH, we have the masks outputted by the segmentation methods. Binary masks classifying those voxels on which the



probability of pertaining to a WMH is above a certain threshold. Note that there is a wide range of evaluation methodologies and coefficients to score segmentation methods, some of which are discussed in Sec.4.1. Nevertheless, we would have to restrict to those which can apply with the output images of our methods.

The general idea behind segmentation masks is to overlay them on the original image to visualize the segmented regions. As explained, this is exactly the kind of masks outputted by the algorithms but it is not what we have as a manual standard. We could have coarse-grained the result masks in unique coordinates and compare them with our manual standard but, we favored the opposite approach. Doing so because it opens the possibility of using the usual evaluation coefficients and is the typical way to proceed. Consequently, we created two different masks from the manual segmentation of WMH.

First, we used the data defining the centers of the WMH to create binary masks. We created these masks by simply generating NifTI binary images in which all the voxels categorized as WMH centers were assigned a 1 while the rest of the voxels were assigned a 0. These masks are quite direct to compute but they do not represent all the information given by the annotations of the expert neuroradiologist.

With the purpose of adding the categorization of sizes given by the expert, we did a second reference standard mask for each subject. To do so, we used the FSL library to apply a Gaussian filter in our binary evaluation masks. The deviation chosen for the Gaussian filters depended on the categorization of each WMH. Then, the deviations for the different Gaussian filters where, 6 mm for *big* ones, 4 mm for *medium* ones, 2 mm for *small* ones and 1 mm for *dot* ones. An example of Gaussian mask overlaid over the respective FLAIR image is shown in Fig.3.4. We hope these Gaussian masks to be a more realistic picture of the WMH positions and shapes.

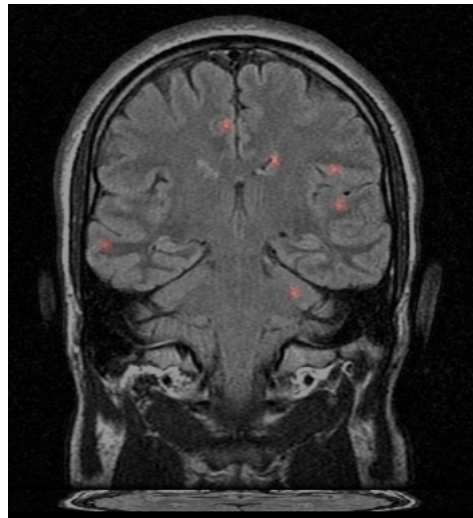


FIGURE 3.4: FLAIR section with the Gaussian reference standard mask overlaid in red.

An important consideration is to understand that WMH are not spherical in general. If we do use Gaussian filters resulting in spherical shapes to define the probability volume of our hyperintensities is because we do not have any data giving the exact shape of the WMH. Using these isotropic shapes introduces a systematic error on the posterior evaluation but, as we are no experts on the topic, we thought this path to be better than segmenting it by ourselves.





## 4 Evaluation

This section is divided in two main parts. First part is dedicated to formally define the evaluation metrics. While doing so, we also explain the different pros and cons of these metrics hence explaining why these metrics are fitted for our case scenario. Special attention is given to explain why these metrics can be useful for a posterior understanding of the methods results. Then, second part is dedicated to display the results of the selected methods for each metric. Moreover, we give some key concepts for a better understanding of why each metric is obtaining the respective performance. To do so, we also present some external experiments done besides the evaluation.

### 4.1 Evaluation Metrics

Here we describe all the evaluations metrics we used in our work. Reading through literature one can find a wide number of different coefficients. Some that usually appear are Dice Similarity Coefficient, F1-score and Hausdorff Distance. In fact, we use two different types of coefficients for this work. Dice Similarity Coefficient, Hausdorff Distance and Average Volume Difference which compute its evaluation taking into account each voxel and on the other hand, Recall and F1-score which evaluate on individuals lesions formed by groups of voxels. The formal definition of each metric and the differences between both types of metrics are further discussed in this section.

In addition, we imply these coefficients to be a good set of evaluation metrics as they supply information about the typical issues of our topic. As explained, they measure differences voxel wise but they also measure WMH as whole entities which is a higher layer of abstraction. These metrics capture the differences in shapes and volumes of WMH which help to understand how neat the results are. Moreover, they are all typical Computer Vision metrics which are relatively safe to use when facing unbalanced data as is in our case (much more non-WMH, voxels than WMH voxels, see Fig.3.3).

#### 4.1.1 Dice Similarity Coefficient (DSC)

The Dice Similarity Coefficient, DSC for short, is the most common metric through all the literature used for the evaluation of WMH Segmentation. Calguiri et al. in [8] where the first to state the extended use of this metric among WMH segmentation evaluation. They went even further and propose it as the main and essential coefficient for testing the performance of WMH Segmentation methods.

The DSC is not an special metric for this task but a general and well known coefficient to test the similarity of two samples. Mathematically it can be expressed as;

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, \quad (4.1)$$

where  $X$  and  $Y$  are the two sets of elements from which we want to measure the DSC. In our case, the set  $X$  and  $Y$  correspond to all those voxels which are classified as WMH in the results and the reference standard masks respectively. The coefficient has also a binary version which is more useful for us. The binary version is expressed mathematically as;

$$DSC = \frac{2TP}{2TP + FP + FN}, \quad (4.2)$$

where TP are those voxels classified as WMH in both sets, FP are those voxels classified as WMH in the results but not in the reference standard masks and FN are those voxels classified as non-WMH, in the results but not in the reference standard masks.

Note that the TN classifications don't appear in Eq.4.2. The goal of the present methods is to detect WMH in human brains. A key point for this relates to the structure of human brains. In human brains we find that the proportion of WMH is small compared to the proportion of other tissues as normal white matter or grey matter. Using a metric which positively scores the correctly classified non-WMH voxels would add a lot of redundancy to the results. Then, using DSC seems a reasonable choice as it doesn't account for the TN classifications and enhances the TP classifications.

#### 4.1.2 Hausdorff Distance

The Hausdorff Distance or HDF for short is another well-known metric for comparing distances between two sets. It measures the maximum distance from all points in one of the sets to the closest point in the other set, see Fig.4.1.

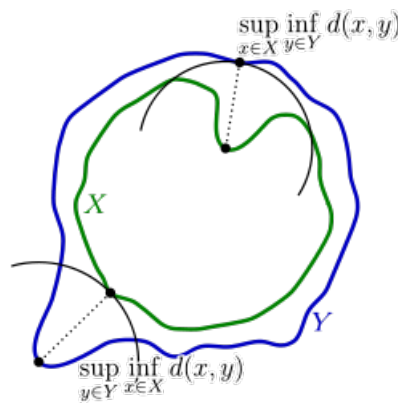


FIGURE 4.1: Representation of the two possible candidates for the Hausdorff distance of two given sets  $X$  and  $Y$ .

The Hausdorff Distance is usually considered as a more descriptive metric for the distance between two sets than the typical minimum distance. Moreover, it also displays interesting outputs when applied to overlapping sets as it happens in some of our results.

As explained the Hausdorff Distance is measured over two sets, in the present work, the two sets refer to the whole group of voxels classified as WMH,  $X$  in the results masks and  $Y$  in the standard reference masks. This implies that only one distance is obtained for each subject of our data. Using our two defined sets  $X, Y$  the Hausdorff Distance can be expressed mathematically as;

$$HDF(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\} \quad (4.3)$$

where  $d(\cdot, \cdot)$  refers to the Euclidean distance between two points. The Hausdorff distance is not as widely used as the DSC coefficient for WMH Segmentation evaluation but it is a common measure in Computer Vision which is the discipline in machine learning to which our methods pertain.

### 4.1.3 Average Volume Difference

The Average Volume Difference or AVD for short is, as its names explains, the average difference between two volume sets. Again, this metric only uses the voxels classified as WMH by the result masks or the reference standard masks not accounting for the non-WMH, voxels in both masks.

The metric must be interpreted carefully as it only compares the sum of every WMH voxel of each mask. As an example, it could be possible that no WMH is correctly classified by a method for a given subject but that the metric scored a perfect score for the respective subject. As a consequence, this metric is only used to analyze how different the result volumes are in respect to the standard reference ones. We expect this average difference to peek if the methods are producing rather big or small WMH volumes in comparison to the real ones more than checking if they are really detecting WMH.

Note that this metrics range is not fitted because, although not happening in our data, the volume of WMH in the reference standard masks could be null. Anyway, there are some of the subjects data which have relative small volumes of WMH making possible big values of the metric. The metric can be written mathematically as;

$$AVD = \left| \frac{\sum_{x \in X} x - \sum_{y \in Y} y}{\sum_{y \in Y} y} \right| \quad (4.4)$$

where  $X$  and  $Y$  are the subsets of classified WMH in the result and reference standard masks respectively.

### 4.1.4 Recall and F1-score for individual lesions

Recall also called sensitivity and F1-score are two typical metrics used in Machine Learning for the evaluation and training of models. It is usual to test recall and precision instead of the F1-score but as we commented before, our data is extremely unbalanced hence replacing precision by F1-score is the proper choice. One could argue to add a higher F1-score but we do not find it necessary for the present work.

Before entering in the mathematical detail behind these metrics, we must state that these metrics are computed taking into account *whole* hyperintensities. To do so, we took those voxels on the results which were classified as WMH and connect them with the direct neighbours if they were also classified as WMH. We define the neighborhood of one voxel as those voxel in contact with a face, edge or vertex of

the principal voxel. Using this connected components we got a picture of full WMH and not just voxels where there is a WMH. Then, we apply the two metrics for the whole WMH, stating that a WMH is well classified if some of its voxels overlaps with a WMH in the reference standard mask.

Recall is the metric which measures the proportion of correct classification and the total number of possible true classifications. For us, it simply states the proportion of WMH detected. As an example, if on method detects three of four WMH for a given subject, it will score 0.75. Mathematically, the recall can be expressed as;

$$recall = \frac{TP}{TP + FN} \quad (4.5)$$

Then F1-score has the same mathematical expression as in Eq.4.2 but using the connected components as evaluation elements. To keep a clear notation, we write DSC when referring to the metric evaluated on individual voxels and F1-score when evaluated for whole WMH.

Although less direct to interpret, F1-score gives a more realistic picture of the performance of the methods. This is because Recall score benefits from not being punished by false positives. As an extreme example of how this may bias the scores of the metric, we could think of a method scoring all the voxels as WMH, then the Recall score will be 1.0 but the F1 would be extremely low for our unbalanced data.

#### 4.1.5 Final Rank

As a closure metric for evaluation, we added a final ranking metric, FR from now on. This metric ranks the performance of the methods over the rest of metrics. The metric differs from the rest as it does not evaluate the methods performance by itself but relatively to the performance of the other methods in the same metric. This ranking is averaged equitably for all the previous metrics.

More explicitly, to compute the FR all methods are ranked in each metric with a score between 0 and 1. Best method for a given metric is given a 1 while worst method is given a 0. Then, the other teams are ranked between (0,1) relatively to their performance within the range of that metric. Finally, the ranks for each method and metrics are averaged resulting in a score in the range [0,1]. Mathematically, the rank of a team A can be expressed as;

$$FR(A) = \frac{1}{N_M} \sum_i^M \frac{r_i^A - r_i^{min}}{r_i^{max} - r_i^{min}} \quad (4.6)$$

where  $r_i^A$  is the score of the team A for a certain metric and the  $r_i^{min/max}$  are the minimum and maximum scores for that metrics. The set of metrics  $M$  comprehends the 5 different metrics used in this work, DSC, HDF, AVD, F1 and Recall. Consequently,  $N_M$  is equal to five with our metrics selection.

Clarify that this metric is only used with the purpose of comparison between methods as it is related to the overall performance of methods. As an example, we could have a method scoring 0.0 at four of the five metrics and having the best FR score. This being said, the metric is useful to give a final idea of how each method performed against each other.

## 4.2 Evaluation Results

Once given the definitions of all the metrics we move on to give the performance of each method on our data. We also devote some lines to comment on the results and reason the differences between segmentation methods.

All the studied methods have been trained using at least the Dice Similarity Coefficient as a loss function. Despite, the data we are using has some major differences to the data used for training. We expect these structural differences between our data and the training one to reduce the performance of the methods. However, the whole point of the study is precisely to find the most accurate and general methods for WMH segmentation independently of the data they are acting on.

### 4.2.1 Dice Similarity Coefficient (DSC)

As we explained in the definition of the DSC, it doesn't suffer from negatively unbalanced data. Despite, checking Fig. 4.2 we realise that the performance of the methods is really low compared to the results obtained for the WMH Segmentation Challenge. This was the first time we faced the importance of developing generalized methods.

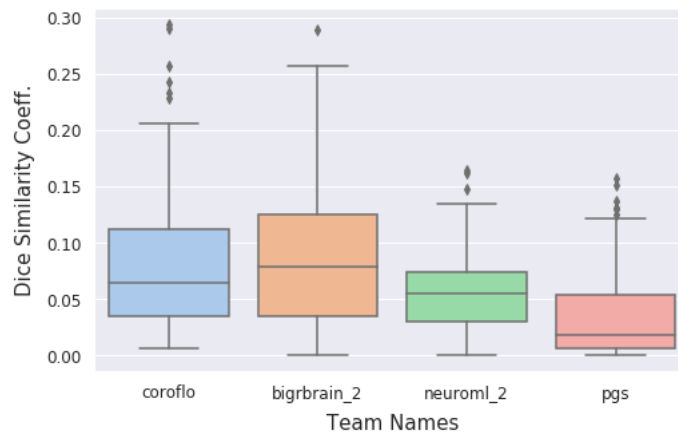


FIGURE 4.2: Box plot for the Dice Similarity Coefficient of the four selected methods.

There are several possibilities to check in order to find why the methods are scoring so low values at DSC. One we expect to happen for all metrics is the differences between the training data and the data for this work. Although both datasets are same types of MRI sequences, the axis in which the sequences were taken and the resolution of them presents strong discrepancies, see Sec. 3.

Optimizing on the DSC, the segmentation methods have *learned* to avoid false positive classifications by increasing the conditions within a voxel would be classified as a WMH. Then, while this has proved useful for achieving high performances on data similar to the original one, it fails in our data. Understand that this is an extremely hard job for the network. To classify WMH for our data, they can't rely on knowledge of raw spatial structures but in more abstract ones.

It is also interesting to note that, although the performance of methods is low, there are several differences between methods. Striking observation is that PGS, which is the clear first classified in the WMH Segmentation Challenge, moved to the last position in our study. This may be consequence of the network topology or that the network is overfitted to the data. Truth is that, we found it to be a mix of both.

The network is overfitted to the data in the sense that it misses when changing its structure but it is not only as consequence of the training but also of the network topology. We see that Coroflo which use MD-GRU instead of U-Net get better performances. In Ref.[19], the authors already realized that MD-GRU generalizes better for different MRI configurations. On the other hand, BigRBrain 2 is a plain U-Net and also gets a good performance due to its CRF reference map.

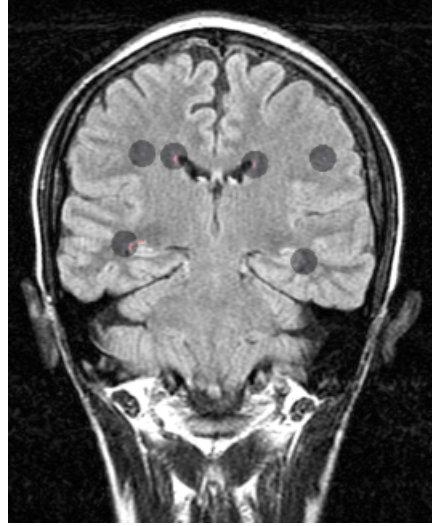


FIGURE 4.3: FLAIR section with reference standard Gaussian masks overlaid in blue and with voxels classified as WMH by Coroflo in red.

Another important limitation for the methods to get good performances is how we defined our reference standard masks. As explained, we did Gaussian masks of different sizes according to the classifications and sizes we were given by the expert neuroradiologist. This sizes were used as deviations for the Gaussian filters which in 3D outputted spherical shapes with decreasing probability. Then, checking Fig.4.3, we can see that while working with binary masks, those masks become simply spheres. This figure also helps to make an important realization, our reference standard lacks precision while defining the WMH. This is really important for the DSC as the methodologies depart from a base in which they would not be able to account the 100% score. The methods have been trained to find all the voxels corresponding to WMH but, in our reference standard masks, the only voxel with a 100% probability of being a WMH is the center one.

#### 4.2.2 Hausdorff Distance

The Hausdorff Distance searches along all the voxels classified as WMH by the methods. The resultant distance given by the metric is, the distance between the furthest positively classified voxel and the closest WMH sphere in the reference standard mask. Note that this only gives one distance for each subject.

Checking Fig.4.4, we see that Coroflo is the method which clearly performs better, followed closely by BigRBrain 2 and NeuroML 2. At the end sits PGS which not only has the worst average but it also has the strongest discrepancies between subjects. The Interquartile Ranges (IQR) for the first three methods are in fact relatively small. Comparing the results with those obtained on the WMH Segmentation

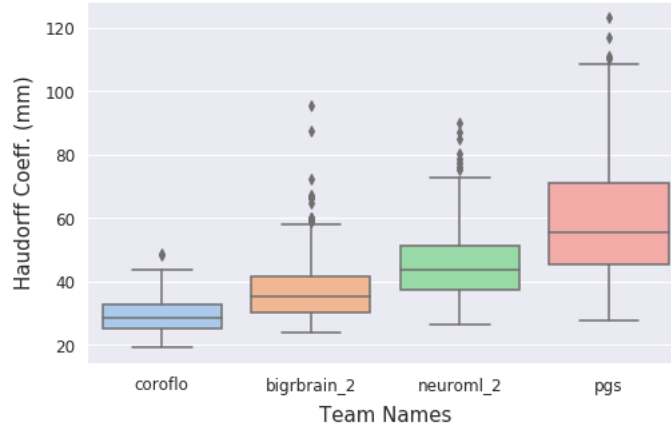


FIGURE 4.4: Box plot for the Hausdorff Distance of the four selected methods.

Challenge, we realize that for this metric the differences in the results are not as significant as for DSC. This is probably because in both scenarios all methods score false positive values determining the final distance for the given subject.

An important realization is that none of the methods have the 6 mm mark (maximum Gaussian masks' standard deviation) in the 1.5 Interquartile Range. This means that in every subject result there is at least one classified WMH that is not inside a sphere of the reference standard mask. In other words, all the results masks contain false positives. Apart from this and the differences in scores between methods, we are not able to output any other enlighten information from this metric. Understand that once the network scores a false positive, the final distance is kind of random.

### 4.2.3 Average Volume Difference

Now, we have already achieved the basic knowledge about the methods performance. Specially, about their capacity to correctly classify voxels as WMH. It has also been detected that some of the classifications are incorrect. Evaluating the methods on the AVD we hope to attain knowledge of the whole volume of voxels classified as WMH. Then, try and check which methods are detecting big or small volumes of WMH to give an idea of how many missclassifications we may have.

To this purpose, we evaluate the methods on the AVD which basically accounts for the relative difference of WMH volume between the reference standard masks and the result masks. Understanding the previous results, we do not expect this metric to be a good metric for evaluation on its own but just a further step. As an example of this metric deficiencies, imagine a subject which has a *big* WMH (6 mm of diameter) and several *dot* WMH (1 mm diameter). In this scenario, the metric might give more relevance to the *big* WMH than to all the other WMH. The truth is that as a matter of practicality, it is usually more interesting to automatically detect small WMH as those are usually the ones hard to find by the expert neuroradiologists.

After studying the scores of the method on the AVD metric, we found some interesting discrepancies for how the methods face the detection problem. NeuroML 2 chooses an aggressive path and the others a more conservative one. When visualizing the result masks outputted by NeuroML 2, we realize that the method has a strong bias to form groups of positively classified voxels. The method is aggressive in the sense that once it detects a WMH voxel, it strongly rises the probability of



their neighbors being also WMH. We do not find this bias so clear in other methods in which some isolated voxels are classified as WMH. As result of this two different approaches or was of training, NeuroML 2 outputs much bigger WMH volumes in most cases.

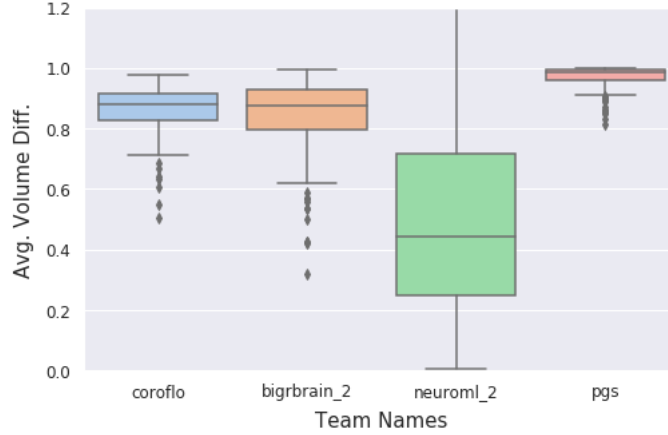


FIGURE 4.5: Box plot for the Average Volume Difference of the four evaluated methods.

Checking Fig. 4.5, we see that NeuroML 2 does beat the rest of the methods for this metric. We also see that its IQR ranges from 0 to 1.36 times the original volume which displays a strong inconsistency between subjects. NeuroML 2 does get big volumes or not depending on how many WMH detects. For some of the subjects, the method was able to detect up to 23 WMH resulting in a big final volume. For others, it simply didn't detect any hence having a null final volume.

The performance of Coroflo, BigRBrain 2 and PGS in this metric is quite similar. They all lay around the 0.9 mark with small IQRs and having some outliers. This indicates that the total number of WMH voxels in the result masks is roughly half of the total number in the reference standard masks. Important to comprehend that the metric does not count if a positive classification is true. As an example to clarify this, it could happen that a method scored a 0.0 at AVD while only scoring false positives.

After going through all the results, we found that  $\sum_{x \in X} x$  is always smaller than  $\sum_{y \in Y} y$ . This means that the reference standard masks always contain bigger volumes of WMH than the result masks. Again, we attribute this difference to the definition of our reference standard mask. The rough classification of sizes along with the non-specification of shapes, results in many voxels classified as WMH while they are not. We could have used smaller Gaussian masks but then, the methods would have it harder to obtain good performances in the individual lesion metrics. In our understanding, those are the most relevant metrics as they can be easily translated to the utility of a metric to correctly detect WMH.

#### 4.2.4 Recall for individual lesions

Until now, the three metrics computed refer to individual voxels or the set of all voxels classified as a WMH. On the other side lay our Recall and F1-score which are evaluated on WMH entities. As explained in the definition of the coefficients, this scores take into account groups of voxels which are connected and classified as an unique WMH. Understand that this is probably the most useful approach because it directly points out where the WMH lay independently of how well they are doing on determining the exact shape.



Team Name	NeuroML 2	Coroflo	BigRBrain 2	PGS
Max. Recall	0.576	0.564	0.436	0.387

TABLE 4.1: Maximum individual lesion Recall score for each method.

Extracting information from this metric is quite straight. Initially, we only had the central voxel of each WMH as reference standard hence the final score for the Recall is basically the proportion of unique WMH detected by a given method. While doing the evaluation, we also saved the maximum proportion of WMH detected by each method as shown in Table 4.1. See that for this metric, although the results look more appealing, they are not comparable to those obtained in the WMH Segmentation Challenge. In the WMH Segmentation Challenge all the Top 10 methods scored above 0.75 at Recall. Despite, for some subjects, NeuroML 2 and Coroflo even detected more than half of the WMH. Those are many WMH detected as no subject had less than 20 and most of them where *dot* size. This confirms that using these methods may be of some use to detect the small WMH which are usually hard to spot by expert neuroradiologists. Of course the detected WMH should be double-checked to assert that they are not a false positives.

The final results for each method are shown in Fig.4.6. In it, we can see that PGS fails again to get a decent score for this metric. This basically leaves the method out of the game as it hasn't been able to give much information about the WMH in any aspect. Coroflo, BigRBrain 2 and NeuroML 2 perform first with significant scores. These relatively high scores are to mention as we must remember that they are being used in data with big differences from the one they were trained at. This implies that the methods had gain some understanding of the structure and intensity distributions that categorize a WMH in a brain MRI independently of the configuration.

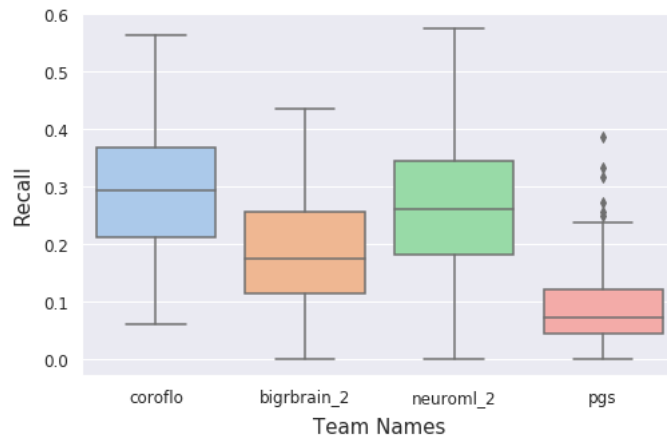


FIGURE 4.6: Box plot for the Recall of the four evaluated methods.

As a final remark, the rather big spheres we created for our reference standard masks have played an important role on the final scores for the metric. The scores while evaluated against the binary masks were all much lower and all the methods were performing like PGS. The methods are not always finding the center of a WMH but they do get close to it. Again, we find this approach the most useful as a first approach.

### 4.2.5 F1-score for individual lesions

F1-score coefficient for individual lesions is the last coefficient we evaluated for each method separately. It is useful to compare the results on this metric to those at the DSC. If this is so is because the actual computation is exactly the same. The only difference being F1 is evaluated on unique WMH and DSC was evaluated voxel wise. Then, we should keep in mind Fig.4.2 while extracting conclusions of the outputs for the F1 metric.

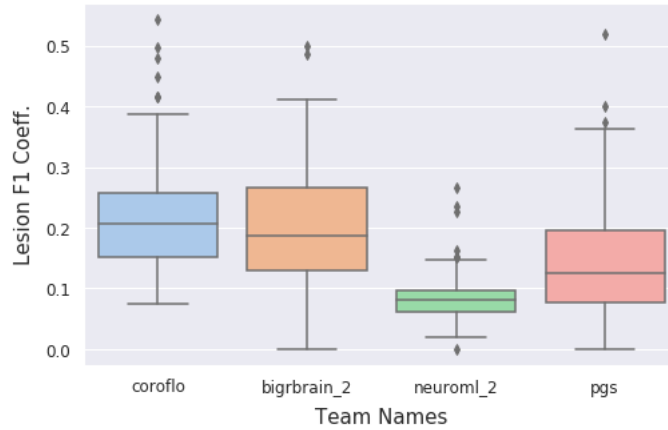


FIGURE 4.7: Box plot for the F1-score of the four evaluated methods.

F1 is a metric designed to penalize the false positive and false negative classifications. As a consequence, having a really unbalanced data as we do, methods which tend to have a low threshold for classifying voxels as WMH would suffer in this metric. This is the case for NeuroML 2 which has a tendency to classify some specific areas of the brain as WMH. This is also related to the big volumes we detected for the method while analysing the AVD metric. In fact, looking at Fig.4.7, we see that here even PGS performs better than NeuroML 2. Still, PGS has a lower performance than Coroflo and BigRBrain 2. Interestingly, these last two methods do get better scores than for the DSC metric but, they do score worse than for the Recall metric. The reading is simple, although the methods are performing reasonably at detecting some of the voxels corresponding to a WMH, they do missclassify many voxels which are not. In other words, in the detected WMH there's some overlap of voxels classified as WMH which makes better results for Recall but, the overlap is not near perfect hence the lower F1-scores. Here again, the imprecision for the shapes of WMH in our reference standard masks play a key role.

### 4.2.6 Final Rank

Finally, we present the results of our Final Rank metric in Fig.4.8. The metric just evaluates all methods against each other using the scores in the previous metrics. This doesn't give any real idea of how the methods are performing on the data by their own but, it helps us *choosing* which would be our main candidate.

First realization we make looking at Fig.4.8 is the low performance of PGS. We already realized and commented while reasoning the metrics that PGS seems to fail at generalizing for our data. We must remind that, before using PGS, we had to modify some of its network parameters as it couldn't fit our data. Apart from this, we did follow the exact procedure while preprocessing the data before fitting it into the

model so, PGS must be scoring these low performances due to its network topology or due its network training.

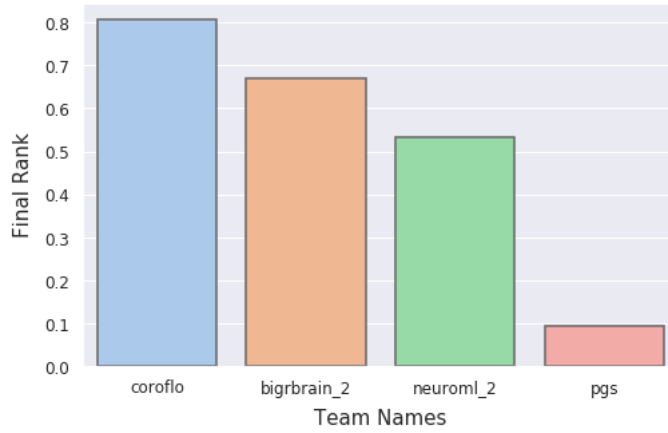


FIGURE 4.8: Box plot for the Final Rank of the four evaluated methods.

More than 0.5 points above PGS sits NeuroML 2. While analysing its results, we saw that this method doesn't generally fail to detect WMH of a subject but it has strong discrepancies between subjects. As an example, it is the method which has the maximum Recall (Table.4.1) but then, it also has the worst score, 0 out of 47 WMH for a subject. Moreover, NeuroML 2 scores are high in part for its score in the AVD which we do not find completely fair. The reason behind using this metric is good enough but, NeuroML 2 approach to classify big WMH volumes plus the deviation in our reference standard mask volumes make its score in the AVD unfair for the rest of methods. Still, it comes up at third position and it does get many good classifications for some subjects. In fact, another good point for NeuroML 2 is that it was the only method directly applicable in our data. NeuroML 2 uses a DeepMedic architecture for its network which has proven useful in many brain segmentation areas and should be further checked for WMH segmentation.

Then at the top positions we have Coroflo with a score of 0.81 followed by BigRBrain 2 with a score of 0.67. This two methods account for good results despite facing really different data and a flawed reference standard as ours. We also want to remember that the original Coroflo was an ensemble of five different architectures based on MD-GRU from which we could only use two. In fact, both methods perform quite similar in all metrics but the HDF and Recall in which Coroflo stands as a winner. Then, to find the maximum number of we should take Coroflo as our first candidate. If our purpose is simply to segment WMH as good as possible, the difference between these two methods is really small. In fact, the better option would be to use both or an ensemble of them.



## 5 Final Discussion

This work includes by performing a systematic review over the WMHs automatic segmentation methods published in peer-reviewed journals, and an objective validation of their accuracy using an independent dataset previously evaluated by an experienced neuroradiologist. The review found a considerable number of methods with a wide range of approaches. While going through those, a judgment was made, the field lacks from standardization as most methods were too data specific, their code was not available, or their evaluation was hardly informative. Mention that while doing the review, some other reviews were gathered where this issue was already stated. Aside from the final selected methods, a section is focused on some methods which may be relevant under certain circumstances but didn't pass the selection criteria. Finally, the only methods that a priori fulfilled all the selection criteria were those presented to the WMH Segmentation Challenge. Then, as the main goal of the platform was to evaluate the current available methods, we took only the ten first methods from the challenge. All those methods were supposed to be freely available, ready to use and MRI independent. After some more research this was found to be false. Only four of the ten methods were practicable for the present study. Despite, the WMH Segmentation Challenge is a useful platform for the evaluation of new methods which fulfill a great need in the field. A point in their favour is that methods can only be submitted as Docker images which makes it easy for external parts to use.

In our opinion, one of the defects of the WMH Segmentation Challenge is that it uses data which has been already preprocessed. It is true that the preprocessing steps executed on the data are few. However, these steps were found to be completely unnecessary. The preprocessing steps are in opposition with the whole purpose of the platform to generalize the evaluation of ready-to-use WMH segmentation methods independently of the data. In fact, some of the libraries used to preprocess the data were already deprecated and while using the methods, it was found that most of them used the raw data instead of the preprocessed one. Moreover, this preprocess steps break with the idea of general ready-to-use methods as posterior users should preprocess their data before feeding it into the methods. Of course, this process is not as hard and as time consuming as coding or training a full method but, it would be better if new methods implemented these preprocess steps by themselves as it would be perfectly possible and easier for users.

An important deficiency of this work is the lack of reference standard masks defined voxel by voxel. Then, while defining the reference standard masks for evaluation, the decisions made were those which favoured the posterior evaluation of individual WMH as a whole and not favouring independent voxel classification. This decision is justified because usually, the exact shape of a WMH is not that important or is work could expert can do once the WMH has been located. In fact, the shapes given for the WMH are always approximations as the resolutions of the MRIs are usually low for at least one of the three spatial axis. So, as a first step, the methods should account for identifying the positions of WMH and so is our definition of reference standard masks. In addition, most of the WMH in our data were classified

as *dot* size meaning that, their actual size was just one voxel. Having a reference standard of that nature would have masked the real findings of the methods.

Once all the methods were evaluated on the defined metrics, some conclusions were made after a deep study of the methods and their resultant masks. First of all, the scores for all the methods fall abruptly in all the evaluations metrics when compared to the scores obtained at the WMH Segmentation Challenge. As the methods are encapsulated in Docker images and the same preprocessing is performed on the data, the only possible element causing these discrepancies is the nature of our data and reference standard. It is much likely a transfer learning problem. Analyzing deeply the possible differences between data the most relevant differences are in the direction in which the scans are made. Usually, the direction in which the scan sequence is made has much less resolution, in fact, this is the case for our data. Then, we find differences between voxel dimensions in our data and the data on the WMH Segmentation Challenge to be the primary suspect for the abrupt decay of performance of the methods. Then, for specific evaluation metrics as the Dice Similarity Coefficient and the Average Volume Difference, the spheres of our reference standard masks introduce a systematic error which further diminish the performance of the methods in these metrics.

After considering the score values of the methods, the relative differences between them was further analyzed to conclude which of the four methods is the best candidate. While PGS was the method with highest performance with the data of the WMH Segmentation Challenge, it yielded lower scores than the other three methods in our dataset. Next method in the ranking is NeuroML 2, its performance in each metric has been already commented and overall, *Neurom 2* does get some good scores and it even gets the highest Recall mark. Its main flaw being that it outputs many false positives even when searching for whole WMH. It is true that gets the best Recall along with Coroflo but it also gets a much lower F1-score displaying a big number of missclassifications. Moreover, it has the issue of overextending the size of each WMH also scoring lower values for voxel wise metrics. Lastly, Coroflo and BigRBrain 2 remain. Coroflo scores 0.14 points more than BigRBrain 2 in the Final Rank metric but the scores for all metrics except Recall are similar. Then, our advice would be to overlay both result masks to account for much better results. In case only one had to be chosen, Coroflo would be our best candidate as it is the one of both with better Recall. Hence, in general, Coroflo would account for more whole WMH than BigRBrain 2. Note that we could use the result masks to locate WMH but we must check for possible false positive values using any of the methods.

In conclusion, we found that the field of WMH automatic segmentation still lacks general and ready-to-use methods. Yet, we also found that, there are some methods which are already getting outstanding performances for some MRI configuration under credible evaluation metrics. In addition, huge advances have been made in the last five years, suggesting that a few years there might be good general and ready-to-use WMH detection methods.

## 6 Further Work

During the time of doing this work, we came up with some interesting ideas which we wanted to test. Some were executed and introduced on the work while others were left out for future works. In this section, we refer to those ideas not carried out.

As explained, the WMH Segmentation Challenge is now an evaluation platform in which many methods are submitted for evaluation. For our work, we only considered the first ten at the date of selection. In our evaluation of those methods we found that the scores inverted over the ones in the Challenge. We do not think this to be general in any sense but it does make a point to do the evaluation on all the methods submitted. This should probably be a continuous work so we contacted the authors of the WMH Segmentation Challenge to add more data and avoiding the need of an external evaluation. Nevertheless, doing the evaluation in all the methods is the most easy way to find more deficiencies of the evaluation platform.

One point we criticised from the WMH Segmentation Challenge is that they pre-processed the data before feeding it into the methods. As commented, this would probably be best if dropped. Nevertheless, one way to check the range of this is to feed the methods with non-preprocessed data and analysing if it is noticeable on the performance of the methods. Note that we did try for some subjects for all methods and the results were nearly the same.

Another important point we wanted to find is a way to further proof the non-generalizability of the methods trained for the WMH Segmentation Challenge. With this in mind, one possible path would be to test the methods on data with the same structure but with different sources than the WMH Segmentation Challenge. Then, accounting a high performance similar to that of the Challenge would further ground the need of generalization for the methods.

As explained, PGS couldn't be used for our data at first hence we had to adapt it. However, while doing so, we didn't retrain the model as we had only the code for production but not the one for training. Then, as PGS was the clear winner at the WMH Segmentation Challenge, we think it would be great to access the whole code of PGS and retrain the model on the original data plus our data. We expect this approach to get a much robust model thus, a better model for production uses.

Finally, our goal was not to build a good segmentation method but to review the ones in the literature. Despite, we found interesting the idea of building an ensemble of Coroflo and BigRBrain 2 as they both get moderate scores but usually in different subjects. Then, building an ensemble and some kind of selection method for the results, we find that the resultant method would increase its evaluation performance substantially.





# A Systematic Review Table

Order PubMed	Title	Dropping Reason
1	Automated White Matter Hyperintensity Segmentation Using Bayesian Model Selection: Assessment and Correlations with Cognitive Change.	NSA
2	White matter hyperintensities increases with traumatic brain injury severity: associations to neuropsychological performance and fatigue.	NTR
3	Aspirin moderates the association between cardiovascular risk, brain white matter hyperintensity total lesion volume and processing speed in normal ageing.	NTR
4	Association between lifetime coffee consumption and late life cerebral white matter hyperintensities in cognitively normal elderly individuals.	NTR
5	Fully Automatic White Matter Hyperintensity Segmentation using U-net and Skip Connection.	SA + T
6	An improved algorithm of white matter hyperintensity detection in elderly adults.	NSA
7	Multi-Disease Segmentation of Gliomas and White Matter Hyperintensities in the BraTS Data Using a 3D Convolutional Neural Network.	NSA
8	Limited One-time Sampling Irregularity Map (LOTS-IM) for Automatic Unsupervised Assessment of White Matter Hyperintensities and Multiple Sclerosis Lesions in Structural Brain Magnetic Resonance Images.	NSA
9	SegAE: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder.	NSA

10	Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset.	SA
11	Global Burden of Small Vessel Disease-Related Brain Changes on MRI Predicts Cognitive and Functional Decline.	NSA
12	White matter hyperintensities and their relationship to cognition: Effects of segmentation algorithm.	NTR
13	Validation and comparison of two automated methods for quantifying brain white matter hyperintensities of presumed vascular origin.	SA
14	Cross-Sectional Association Between Cognitive Frailty and White Matter Hyperintensity Among Memory Clinic Patients.	NTR
15	Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs.	SA
16	White matter hyperintensity burden in patients with ischemic stroke treated with thrombectomy.	NTR
17	Automated lesion segmentation with BIANCA: Impact of population-level features, classification algorithm and locally adaptive thresholding.	NSA
18	Intra-Scanner and Inter-Scanner Reproducibility of Automatic White Matter Hyperintensities Quantification.	NTR
19	Dilated Saliency U-Net for White Matter Hyperintensities Segmentation Using Irregularity Age Map.	SA
20	White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts - The MRI-GENIE study.	NSA
21	Characterization of White Matter Hyperintensities in Large-Scale MRI-Studies.	NTR
22	Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge.	NTR
23	Magnetic Resonance Imaging of Cerebral Small Vessel Disease in Men Living with HIV and HIV-Negative Men Aged 50 and Above.	NTR
24	Brain imaging correlates of mild cognitive impairment and early dementia in patients with type 2 diabetes mellitus.	NTR
25	MRI white matter lesion segmentation using an ensemble of neural networks and over-complete patch-based voting.	NSA

26	White matter hyperintensities are associated with falls in older people with dementia.	NTR
27	Voxel-Wise Logistic Regression and Leave-One-Source-Out Cross Validation for white matter hyperintensity segmentation.	NSA
28	Altered Whole-Brain Structural Covariance of the Hippocampal Subfields in Subcortical Vascular Mild Cognitive Impairment and Amnesic Mild Cognitive Impairment Patients.	NTR
29	The challenge of cerebral magnetic resonance imaging in neonates: A new method using mathematical morphology for the segmentation of structures including diffuse excessive high signal intensities.	NTR
30	DEWS (DEep White matter hyperintensity Segmentation framework): A fully automated pipeline for detecting small deep white matter hyperintensities in migraineurs.	SA
31	Frontal White Matter Hyperintensity Is Associated with Verbal Aggressiveness in Elderly Women with Alzheimer Disease and Amnesic Mild Cognitive Impairment.	NTR
32	Volumetric Distribution of the White Matter Hyper-Intensities in Subject with Mild to Severe Carotid Artery Stenosis: Does the Side Play a Role?	NTR
33	Validation and Optimization of BIANCA for the Segmentation of Extensive White Matter Hyperintensities.	NSA
34	UBO Detector - A cluster-based, fully automated pipeline for extracting white matter hyperintensities.	NSA
35	White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks.	NSA
36	Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology.	NSA
37	Physical Activity and Changes in White Matter Hyperintensities over Three Years.	NTR
38	Association between Red Blood Cells Omega-3 Polyunsaturated Fatty Acids and White Matter Hyperintensities: The MAPT Study.	NTR
39	Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI.	NSA

40	Nonlinear temporal dynamics of cerebral small vessel disease: The RUN DMC study.	NTR
41	Diffusion tensor image segmentation of the cerebrum provides a single measure of cerebral small vessel disease severity related to cognitive change.	NTR
42	Associations between white matter hyperintensities and cognitive decline over three years in non-dementia older adults with memory complaints.	NTR
43	Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging.	NSA
44	White matter hyperintensities are seen only in GRN mutation carriers in the GENFI cohort.	NTR
45	Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease.	NTR
46	Aortic hemodynamics and white matter hyperintensities in normotensive postmenopausal women.	NTR
47	Improved Automatic Segmentation of White Matter Hyperintensities in MRI Based on Multilevel Lesion Features.	NSA
48	Longitudinal segmentation of age-related white matter hyperintensities.	NSA
49	A challenging issue: Detection of white matter hyperintensities in neonatal brain MRI.	NSA
50	Relationship between white matter hyperintensities volume and the circle of Willis configurations in patients with carotid artery pathology.	NTR
51	Impact of frontal white matter hyperintensity on instrumental activities of daily living in elderly women with Alzheimer disease and amnesic mild cognitive impairment.	NTR
52	Reproducible segmentation of white matter hyperintensities using a new statistical definition.	SA
53	Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease.	NSA
54	Automated segmentation reveals silent radiographic progression in adult-onset vanishing white-matter disease.	NTR
55	Supervised learning technique for the automated identification of white matter hyperintensities in traumatic brain injury.	NSA
56	Mental speed is associated with the shape irregularity of white matter MRI hyperintensity load.	NTR

57	Nonnegative matrix factorization and sparse representation for the automated detection of periodic limb movements in sleep.	NTR
58	Compromised Neurocircuitry in Chronic Blast-Related Mild Traumatic Brain Injury.	NTR
59	BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities.	SA
60	The effects of white matter disease on the accuracy of automated segmentation.	NTR
61	Longitudinal patterns of leukoaraiosis and brain atrophy in symptomatic small vessel disease.	NTR
62	White matter hyperintensities and imaging patterns of brain ageing in the general population.	NTR
63	Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke.	NTR
64	Characterising the grey matter correlates of leukoaraiosis in cerebral small vessel disease.	NTR
65	Subclinical cerebrovascular disease inversely associates with learning ability: The NOMAS.	NTR
66	Automated removal of spurious intermediate cerebral blood flow volumes improves image quality among older patients: A clinical arterial spin labeling investigation.	NTR
67	Effects of vascular risk factors and ApoE- $\epsilon$ 4 on white matter integrity and cognitive decline.	NTR
68	Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI.	NTR
69	Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review.	NTR
70	Frontal white matter hyperintensity predicts lower urinary tract dysfunction in older adults with amnesic mild cognitive impairment and Alzheimer's disease.	NTR
71	Automatic segmentation and volumetric quantification of white matter hyperintensities on fluid-attenuated inversion recovery images using the extreme value distribution.	NSA
72	Longitudinal relaxographic imaging of white matter hyperintensities in the elderly.	NTR

73	Progression of white matter hyperintensities of presumed vascular origin increases the risk of falls in older people.	NTR
74	Automatic segmentation and quantitative analysis of white matter hyperintensities on FLAIR images using trimmed-likelihood estimator.	NSA
75	Automated White Matter Hyperintensity Detection in Multiple Sclerosis Using 3D T2 FLAIR.	NSA
76	Automated segmentation and quantification of white matter hyperintensities in acute ischemic stroke patients with cerebral infarction.	NSA
77	Metabolic syndrome, prediabetes, and brain abnormalities on mri in patients with manifest arterial disease: the SMART-MR study.	NTR
78	Lesion segmentation from multimodal MRI using random forest following ischemic stroke.	NTR
79	Sub-cortical infarcts and the risk of falls in older people: combined results of TASCOG and Sydney MAS studies.	NTR
80	Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images.	NSA
81	White matter hyperintensities segmentation: a new semi-automated method.	NSA
82	Do cardiovascular risk factors explain the link between white matter hyperintensities and brain volumes in old age? A population-based study.	NTR
83	Cerebral small vessel disease affects white matter microstructure in mild cognitive impairment.	NTR
84	[Age-related white matter lesions (leukoaraiosis): an update].	NTR
85	Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images.	NSA
86	Perinatal factors and regional brain volume abnormalities at term in a cohort of extremely low birth weight infants.	NTR
87	White matter hyperintensities, exercise, and improvement in gait speed: does type of gait rehabilitation matter?	NTR
88	White matter hyperintensity burden and disability in older adults: is chronic pain a contributor?	NTR

89	Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review.	NTR
90	Thrombogenic microvesicles and white matter hyperintensities in postmenopausal women.	NTR
91	Most edges in Markov random fields for white matter hyperintensity segmentation are worthless.	NTR
92	Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation.	SA
93	Automatic segmentation of white matter hyperintensities by an extended FitzHugh & Nagumo reaction diffusion model.	NSA
94	Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts.	NTR
95	Brain tissue volumes in the general population of the elderly: the AGES-Reykjavik study.	NTR
96	Average daily blood pressure, not office blood pressure, is associated with progression of cerebrovascular disease and cognitive decline in older people.	NTR
97	Validation of automated white matter hyperintensity segmentation.	NTR
98	Quantitative approaches for assessment of white matter hyperintensities in elderly populations.	NSA
99	Elastic registration of multimodal prostate MRI and histology via multiattribute combined mutual information.	NTR
100	A comparison of different automated methods for the detection of white matter lesions in MRI data.	NSA
101	MRI markers of small vessel disease in lobar and deep hemispheric intracerebral hemorrhage.	NTR
102	Differential patterns of cognitive decline in anterior and posterior white matter hyperintensity progression.	NTR
103	Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T.	SA
104	White matter hyperintensities predict functional decline in voiding, mobility, and cognition in older adults.	NTR
105	Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging.	NTR

106	Ventricular dilation: association with gait and cognition.	NTR
107	Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR.	NSA
108	Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities.	NSA
109	Metabolic risks, white matter hyperintensities, and arterial stiffness in high-functioning healthy adults.	NTR
110	Longitudinal follow-up of individual white matter hyperintensities in a large cohort of elderly.	NTR
111	Three-dimensional MRI analysis of individual volume of Lacunes in CADASIL.	NTR
112	Diabetes increases atrophy and vascular lesions on brain MRI in patients with symptomatic arterial disease.	NTR
113	Misclassified tissue volumes in Alzheimer disease patients with white matter hyperintensities: importance of lesion segmentation procedures for volumetric analysis.	NTR
114	Regional white matter hyperintensity burden in automated segmentation distinguishes late-life depressed subjects from comparison subjects matched for vascular risk factors.	NTR
115	The brain-derived neurotrophic factor VAL66MET polymorphism and cerebral white matter hyperintensities in late-life depression.	NTR
116	Automated and visual scoring methods of cerebral white matter hyperintensities: relation with age and cognitive function.	NSA
117	An automated procedure for the assessment of white matter hyperintensities by multi-spectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases.	NSA
118	Verbal working memory and atherosclerosis in patients with cardiovascular disease: an fMRI study.	NTR
119	Vascular risk factors and white matter hyperintensities in patients with amnesic mild cognitive impairment.	NTR
120	A fully automated method for quantifying and localizing white matter hyperintensities on MR images.	SA



121	Weekly alcohol consumption, brain atrophy, and white matter hyperintensities in a community-based sample aged 60 to 64 years.	NTR
122	Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: structural MR imaging changes and apolipoprotein E genotype.	NTR
123	Fully automatic segmentation of white matter hyperintensities in MR images of the elderly.	NSA
124	White matter hyperintensity progression and late-life depression outcomes.	NTR
125	Evidence of subtle gray-matter pathologic changes in healthy elderly individuals with nonspecific white-matter hyperintensities.	NTR
126	White matter changes in normal pressure hydrocephalus and Binswanger disease: specificity, predictive value and correlations to axonal degeneration and demyelination.	NTR
127	A new rapid landmark-based regional MRI segmentation method of the brain.	NSA
128	Evidence for genetic variance in white matter hyperintensity volume in normal elderly male twins.	NTR

TABLE A.1: Table with all the publication selected after the search on PubMed. The acronyms for Dropping Reasons columns are; no topic related (NTR), no software available (NSA), software available but not trained (SA), software and training available (SA + T).



# Bibliography

- [1] Joanna M. Wardlaw, Maria C. Valdés Hernández, and Susana Muñoz-Maniega. “What are White Matter Hyperintensities Made of?” In: *Journal of the American Heart Association* 4.6 (June 2015). DOI: [10.1161/jaha.114.001140](https://doi.org/10.1161/jaha.114.001140). URL: <https://doi.org/10.1161/jaha.114.001140>.
- [2] S. Debette and H. S. Markus. “The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis”. In: *BMJ* 341.jul26 1 (July 2010), pp. c3666–c3666. DOI: [10.1136/bmj.c3666](https://doi.org/10.1136/bmj.c3666). URL: <https://doi.org/10.1136/bmj.c3666>.
- [3] Mohamad Habes et al. “White matter hyperintensities and imaging patterns of brain ageing in the general population”. In: *Brain* 139.4 (Feb. 2016), pp. 1164–1179. DOI: [10.1093/brain/aww008](https://doi.org/10.1093/brain/aww008). URL: <https://doi.org/10.1093/brain/aww008>.
- [4] Alan J. Thomas et al. “Pathologies and Pathological Mechanisms for White Matter Hyperintensities in Depression”. In: *Annals of the New York Academy of Sciences* 977.1 (Nov. 2002), pp. 333–339. DOI: [10.1111/j.1749-6632.2002.tb04835.x](https://doi.org/10.1111/j.1749-6632.2002.tb04835.x). URL: <https://doi.org/10.1111/j.1749-6632.2002.tb04835.x>.
- [5] Leonardo Pantoni. “Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges”. In: *The Lancet Neurology* 9.7 (July 2010), pp. 689–701. DOI: [10.1016/S1474-4422\(10\)70104-6](https://doi.org/10.1016/S1474-4422(10)70104-6). URL: [https://doi.org/10.1016/S1474-4422\(10\)70104-6](https://doi.org/10.1016/S1474-4422(10)70104-6).
- [6] Matthew J. Kempton et al. “Meta-analysis, Database, and Meta-regression of 98 Structural Imaging Studies in Bipolar Disorder”. In: *Archives of General Psychiatry* 65.9 (Sept. 2008), p. 1017. DOI: [10.1001/archpsyc.65.9.1017](https://doi.org/10.1001/archpsyc.65.9.1017). URL: <https://doi.org/10.1001/archpsyc.65.9.1017>.
- [7] P. Videbech. “MRI findings in patients with affective disorder: a meta-analysis”. In: *Acta Psychiatrica Scandinavica* 96.3 (Sept. 1997), pp. 157–168. DOI: [10.1111/j.1600-0447.1997.tb10146.x](https://doi.org/10.1111/j.1600-0447.1997.tb10146.x). URL: <https://doi.org/10.1111/j.1600-0447.1997.tb10146.x>.
- [8] Maria Eugenia Caligiuri et al. “Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review”. In: *Neuroinformatics* 13.3 (Feb. 2015), pp. 261–276. DOI: [10.1007/s12021-015-9260-y](https://doi.org/10.1007/s12021-015-9260-y). URL: <https://doi.org/10.1007/s12021-015-9260-y>.
- [9] Nikhil Buduma. *Fundamentals of Deep Learning: Designing Next-Generation Artificial Intelligence Algorithms*. 1st ed. O’Reilly Media, 2015. ISBN: 9781491925614.
- [10] Hugo J. Kuijf et al. “Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge”. In: *IEEE Transactions on Medical Imaging* 38.11 (Nov. 2019), pp. 2556–2568. DOI: [10.1109/tmi.2019.2905770](https://doi.org/10.1109/tmi.2019.2905770). URL: <https://doi.org/10.1109/tmi.2019.2905770>.

- [11] L. Shamseer et al. "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation". In: *BMJ* 349.jan02 1 (Jan. 2015), g7647–g7647. DOI: [10.1136/bmj.g7647](https://doi.org/10.1136/bmj.g7647). URL: <https://doi.org/10.1136/bmj.g7647>.
- [12] Ludovica Griffanti et al. "BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities". In: *NeuroImage* 141 (Nov. 2016), pp. 191–205. DOI: [10.1016/j.neuroimage.2016.07.018](https://doi.org/10.1016/j.neuroimage.2016.07.018). URL: <https://doi.org/10.1016/j.neuroimage.2016.07.018>.
- [13] Soheil Damangir et al. "Reproducible segmentation of white matter hyperintensities using a new statistical definition". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 30.3 (Dec. 2016), pp. 227–237. DOI: [10.1007/s10334-016-0599-3](https://doi.org/10.1007/s10334-016-0599-3). URL: <https://doi.org/10.1007/s10334-016-0599-3>.
- [14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. eprint: [arXiv:1505.04597](https://arxiv.org/abs/1505.04597).
- [15] Jisu Hong et al. "Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs". In: *Computer Methods and Programs in Biomedicine* 183 (Jan. 2020), p. 105065. DOI: [10.1016/j.cmpb.2019.105065](https://doi.org/10.1016/j.cmpb.2019.105065). URL: <https://doi.org/10.1016/j.cmpb.2019.105065>.
- [16] Yunhee Jeong et al. "Dilated Saliency U-Net for White Matter Hyperintensities Segmentation Using Irregularity Age Map". In: *Frontiers in Aging Neuroscience* 11 (June 2019). DOI: [10.3389/fnagi.2019.00150](https://doi.org/10.3389/fnagi.2019.00150). URL: <https://doi.org/10.3389/fnagi.2019.00150>.
- [17] Jong-Min Lee Gilsoon Park Jinwoo Hong. "White matter hyperintensities segmentation using UNet with highlighted foreground". In: (2019).
- [18] J. E. Iglesias et al. "Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods". In: *IEEE Transactions on Medical Imaging* 30.9 (Sept. 2011), pp. 1617–1634. DOI: [10.1109/tmi.2011.2138152](https://doi.org/10.1109/tmi.2011.2138152). URL: <https://doi.org/10.1109/tmi.2011.2138152>.
- [19] Marleen de Bruijne Robin Camarasa Corentin Doue and Florian Dubost. "Segmentation of White Matter Hyperintensities with an Ensemble of Multi-Dimensional Convolutional Gated Recurrent Units". In: (2018).
- [20] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. eprint: [arXiv:1406.1078](https://arxiv.org/abs/1406.1078).
- [21] Simon Andermatt, Simon Pezold, and Philippe Cattin. "Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data". In: *Deep Learning and Data Labeling for Medical Applications*. Springer International Publishing, 2016, pp. 142–151. DOI: [10.1007/978-3-319-46976-8\\_15](https://doi.org/10.1007/978-3-319-46976-8_15). URL: [https://doi.org/10.1007/978-3-319-46976-8\\_15](https://doi.org/10.1007/978-3-319-46976-8_15).
- [22] Boris Shirokikh and Mikhail Belyaev. "WMH Segmentation Using an Adjusted DeepMedic Architecture and an Improved Learning Approach". In: (2019).
- [23] Konstantinos Kamnitsas et al. "DeepMedic for Brain Tumor Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2016, pp. 138–149. DOI: [10.1007/978-3-319-55524-9\\_14](https://doi.org/10.1007/978-3-319-55524-9_14). URL: [https://doi.org/10.1007/978-3-319-55524-9\\_14](https://doi.org/10.1007/978-3-319-55524-9_14).
- [24] Marleen de Bruijne Shuai Chen. "An End-to-end Approach with CNN and Posterior-CRF in White Matter Hyperintensities Segmentation". In: (2019).

- [25] Stephen M. Smith et al. "Advances in functional and structural MR image analysis and implementation as FSL". In: *NeuroImage* 23 (Jan. 2004), S208–S219. DOI: [10.1016/j.neuroimage.2004.07.051](https://doi.org/10.1016/j.neuroimage.2004.07.051). URL: <https://doi.org/10.1016/j.neuroimage.2004.07.051>.
- [26] S. Klein et al. "elastix: A Toolbox for Intensity-Based Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 29.1 (Jan. 2010), pp. 196–205. DOI: [10.1109/tmi.2009.2035616](https://doi.org/10.1109/tmi.2009.2035616). URL: <https://doi.org/10.1109/tmi.2009.2035616>.
- [27] John Ashburner and Karl J. Friston. "Voxel-Based Morphometry—The Methods". In: *NeuroImage* 11.6 (June 2000), pp. 805–821. DOI: [10.1006/nimg.2000.0582](https://doi.org/10.1006/nimg.2000.0582). URL: <https://doi.org/10.1006/nimg.2000.0582>.